CSCI 620 – Intro to Big Data

Assignment #6 – Data Cleaning

Name: *Yuvraj Singh*

1) Information Description

A json file is supplied for this assignment. The file represents extra data for certain movies based on movie_ids and movie_titles. The extra data consists of the currency type for the movie, the movie title, the movie id, the cost of the movie, the distributor, and the box office revenue. Each new line represents a new set of information, whereas each line may or may not have some new information for a specific key. For example, a line may have the cost and distributor of the movie, but it may not have the box office revenue.

2) Adding the New Information to MongoDB via Title ID

Since each new data entry (line) in the json file represents a movie, the fastest way to update the movies would be by using the title id in the new data entry since it corresponds to a title in the MongoDB database. The incoming dataset was saving to a Pandas DataFrame. Only rows with title ids were kept and only the first duplicates were kept. After the filtering was done, each movie that had valid data to be updated was processed and updated via the *def get_new_data_via_id.* When the function is ran, the console prints out how many successful updates it had along with some other information that can be seen below.

```
There are 180039 data entries that have a title id associated with them. These entries may or may not have any
valid data to update.
There are 48859 data entries that do not have a title id associated with them. The movies that correspond to
these entries in the extra-data will not be updated.
Out of 180039 entries that do have valid title ids, 132297 ids do not have any new data to update.
Data entries that do not have any valid data to be updated have been removed. The total number of documents
that have data to be updated is 47742
Had to update 47742 documents. Was able to successfully update 47742 documents. There were 0
unsuccessful updates
It took 13.2692506 seconds to update records via id
```

3) Adding the New Information to MongoDB via Title

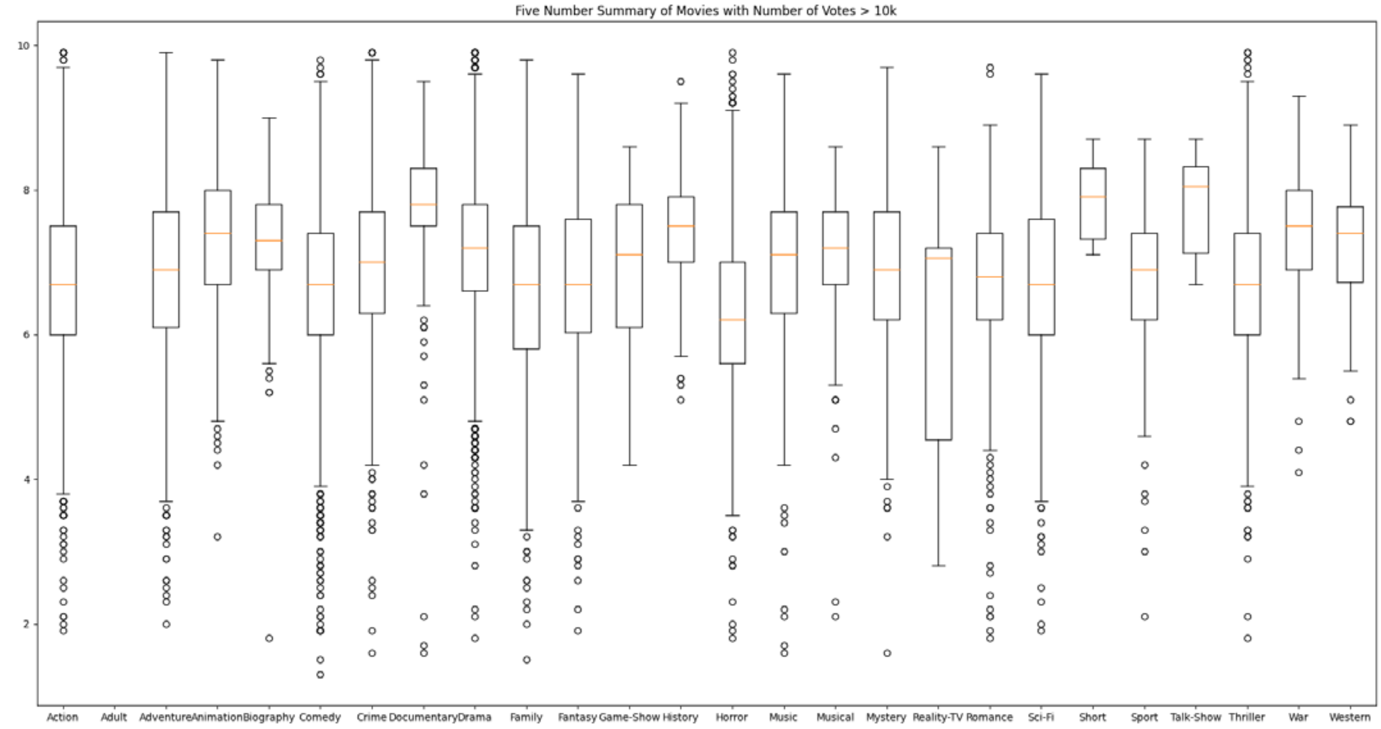
The previous section updated titles in the database by utilized the title ids in the data entries of the extra-data json file. The issue that came up was that some of the entries did not have a title id, only a title. Therefore, a separate method was created to update records via title instead, *def get_new_data_via_title.*

Some of the issues that occur using this method occur while also using the via id method such as duplicates, entries without any valid data to update, entries without titles, etc… Using this method, we notice that there are less documents that can be updated based on using data that only has titles.

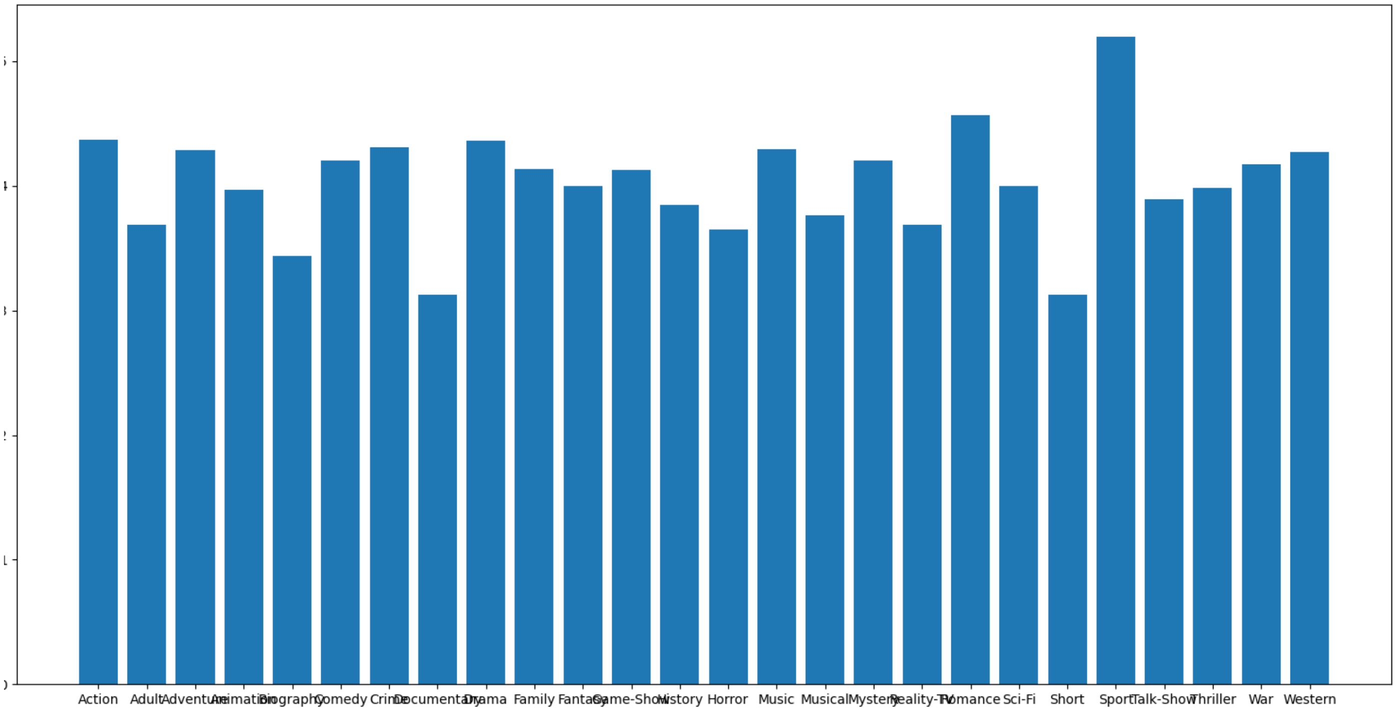
```
There are 114062 new data entries that have titles. These entries can be updated
There are 116763 new data entries that do not have any title associated with them. These entries will not be
able to update any documents
There are 5692 duplicate new data entries
Out of 106211 entries that do have valid titles, 68222 titles do not have any new data to update.
Data entries that do not have any valid data to be updated have been removed. The total number of documents
that have data to be updated is 37989
Had to update 37989 documents. Was able to successfully update 37989 documents. There were 0
unsuccessful updates
It took 19865.6730849 seconds to update the documents that have ids
```

4) Plotting Custom Queries

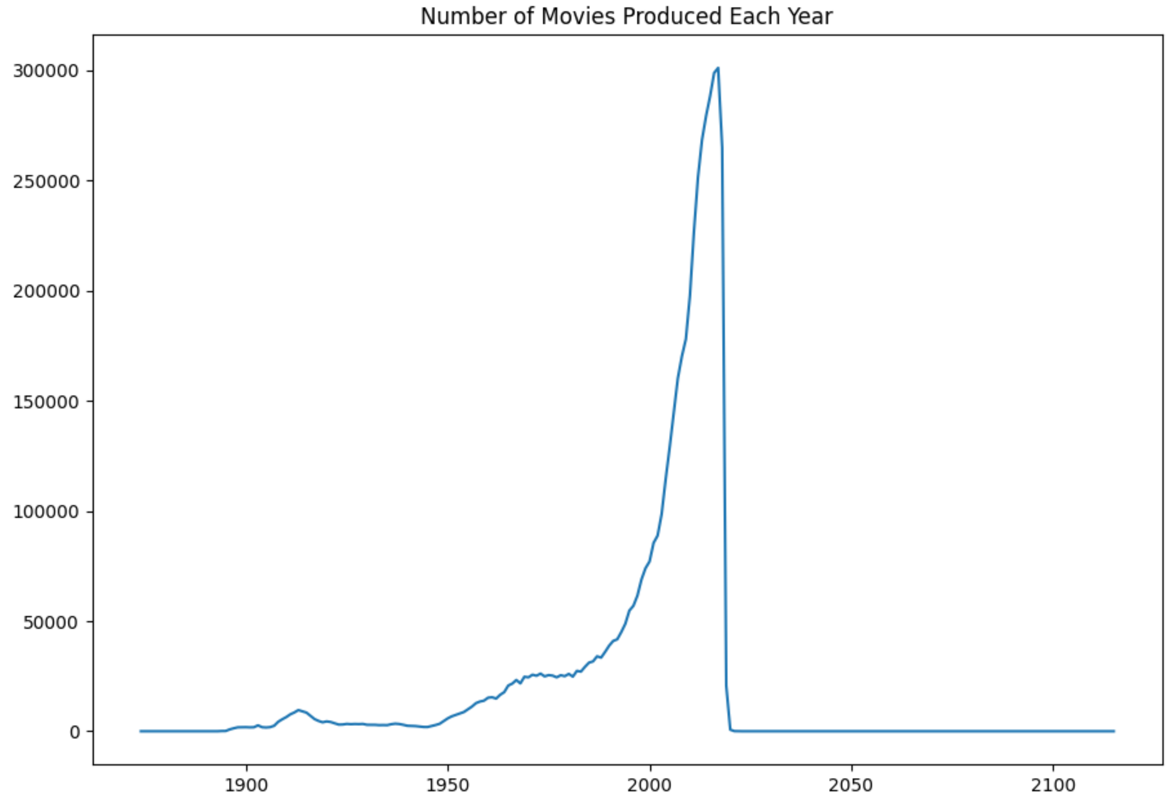
In this assignment, a few custom queries had to be performed and the data had to be categorized as specific charts. The *def query_plots* method is to be run to generate the plots for each graph. For example, the first query had to perform a five-number summary on the average ratings of movies with more than 10K votes for each genre. To do so, the



For the second query, the average number of actors per movie had to be calculated for each genre and was to be plotted as a bar chart.



For the third query, the number of movies produced each year is to be plotted as a time series chart.


Number of Movies Produced Each Year